Synopsis

Proteins, which regulate most of the biological activities, perform their function through their unique three-dimensional structures. It is now established that the millions of protein sequences that exist in nature fold to one of the structures from a limited menu of less than two thousands (Chothia, 1992). By noting that the protein structures are mostly conserved while sequences are more tolerant to mutations, it can be inferred that the native state topology of proteins is encoded with valuable information. The protein topology information can be exploited in order to address the inverse protein folding problem, which is to design sequences for a given structure. In the first part of this thesis, this concept is developed and tested by investigating topologies ranging from lattice models to real protein structures and generating the sequences consisting of monomers ranging from the simplest HP model, monomers of five types to the realistic case of all twenty amino acids.

The second part of the thesis deals with developing refined scoring functions which accurately describe the non-covalent interactions in protein structures. The stability of the structure and functional activities of the proteins depend on the interactions that amino acids residues have among themselves and also with other biologically active molecules. These physical interactions are measured by energy functions. The *ab initio* energy calculations are still not practical because of extensive calculations that are beyond even today's computational power. Fortunately, a large number of X-ray structures are available from which knowledge based statistical energy functions can be derived. Such functions are extremely useful in addressing problems such as structure-prediction, docking and understanding the effects of mutation. Further, it is observed that the local environment of amino acids plays a crucial role in the protein folding, protein–

protein or protein–ligand interactions. For example, protein surface, interior and protein–protein interface are packed in different ways and the propensity of the same amino acid pair can be different depending on the location (e.g., either exposed or buried) of the amino acids. Several potential energy functions and corresponding matrices, which have been shown to be useful in studies such as structure prediction, have been reported in the literature (Li and Liang, 2007). These are derived both at a coarse grained residue level as well as at the atomic level. The previously developed interaction matrices mainly consider the number of interacting pairs of amino acids and amino acid composition in the selected dataset of proteins, but serious efforts to discriminate the interactions in varying environments as described above were not made. We have taken a step in this direction and considered two different types of protein internal environments for the construction of such knowledge-based potential energy matrices. One of these is the local degree (number of contacts) and the other is the secondary structural element of amino acids. We have also revisited the definition of hydrophobicity based on the propensities of amino acids in these varied environments. The investigations have shown that the environment-based interaction preferences for amino acids is able to provide good scoring functions which perform exceedingly well in discriminating the native structure from the structures with random interactions.

Further, the membrane proteins are located in a completely different physico-chemical environment. They have different amino acid composition and also they fold differently. There is a need for a good scoring function to describe interactions in membrane proteins. The availability of substantial number of membrane protein structures, especially the alpha-helical structures motivated us to develop

a much needed scoring function for trans-membrane alpha helical proteins. A detailed investigation of this problem is presented in this thesis.

Apart from developing scoring functions, the packing of helices in membrane proteins is investigated by an approach based on the local backbone geometry and side chain atom-atom contacts of amino acids.

The thesis consists of nine chapters. The first chapter gives a brief background about the work reported in the literature related to the theme of the thesis such as inverse protein folding, topology or protein structure, sequence design, knowledge based potential energy functions, local environment of proteins and helix packing in membrane proteins. It also describes the objective of the work presented in the thesis.

Chapter 2 contains various methods and techniques developed or used in the thesis. Protein structures can be represented as a graph where amino acids are nodes and non-covalent interaction between amino acids are designated as edges. The number of edges on a node (i.e., an amino acid residue) is called the degree of that node. These graphs have been used to derive the topological indices of the protein structure by using the primary and secondary degree of amino acid residues. Using the topological indices of a given structure and grouping of amino acids into five types, we generate optimized amino acid sequences for that specific structure. Further development of environment potential energy functions, and normalization factors, which can be used for calculating the energy for the proteins are discussed. All these methods are discussed in more detail in their respective chapters of the thesis.

Chapters 3, 4 and 5 deal with the studies on topology-based sequence design for given protein conformations. Chapter 3 brings out the importance of the native

state topology in sequence design. An efficient and computationally fast method is presented for ranking the residue sites (i.e., nodes) in a given native-state structure, which enables us to design sequences for a given structure. Ranking of nodes in the graph is done by assigning the judiciously chosen node weights which are based on secondary connections, along with primary connections of the nodes. Three dimensional lattice model wherein nodes are occupied by either H (hydrophobic) or P (non-hydrophobic) monomers is chosen as an example (for which exhaustive enumeration of sequences with HP monomers is practical) to validate the scheme of ranking the nodes. It has been shown that this scheme is able to identify the sequence with lowest energy for all possible composition of H–P monomers in selected thirteen different three dimensional lattice model structures. Further, the real sequences of a few structures are converted to HP sequence. Our scheme is able to predict the protein sequence (with HP monomers) with energy better than the one obtained for the native sequence. Further employment of deterministic optimization technique on these designed sequences improves the results. In summary, the proposed scheme of deriving topological indices is able to design a lower energy sequence for any given structure (Jha *et al.*, 2007).

Nature has selected twenty amino acids as the building blocks of proteins. However experimental and theoretical studies have suggested that a smaller number of amino acid alphabet, grouped on the basis of their physical or chemical properties, are able to capture the gross-essential features of protein structure and folding. This level of simplification facilitates in-depth computational studies. Chapter 4 gives a brief background about the reported sets of reduced amino acid alphabet in the literature. It also gives a set of reduced alphabets obtained from the Multi Dimensional Scaling (MDS) technique. This chapter introduces a systematic

method to compare the available set of reduced alphabets, in which a potential energy matrix of size 20x20, reported in literature for interactions among 20 amino acids, is used as a starting point to obtain the interaction energy values between different monomers of reduced alphabet. The chapter as a whole deals with the computation of inter-residue interaction energies for the reduced amino acid alphabet, which helps in the selection of a better grouping (Luthra *et al.*, 2007).

Since the number of protein folds is limited in the structure space, a large number of sequences can adopt the same fold. However a complete search in the sequence space is not possible because the number of possible sequences for even a small protein is an astronomical number. Thus, the goal is to search energetically minimized sequences in the sequence space for a target conformation. Such a study of designing realistic sequences to known protein structures is explored in Chapter 5 by combining the two techniques developed in the Chapters 3 and 4. Here, we adopt a strategy of generating a sequence of five types of monomers, which are placed at residue sites according to their node weights generated from the topology of the target structure. Further, the five monomer types are later expanded to twenty types based on the grouping. Such a procedure provides a large number of sequences of fixed amino acid composition, which have the same energy according the 5x5 interaction energy matrix, but differ in their energies when evaluated from a 20x20 matrix. The energies of these sequences have been calculated by standard scoring function such as Miyazawa-Jernigan (MJ) potential energy matrix. Further, three different sets of sequences with some constraints such as fixing the positions of structurally or functionally important amino acid residue in the generated sequences have been enumerated. The performance of the designed sequences was tested against control sets of about 100 million random sequences generated under similar constraints. The validation for this

scheme of designing sequences is done by comparing the energy spectrum of the designed and random sequences. The designed sequences turned out to be energetically better than not only the random sequences but also than the native protein sequence. This result is further confirmed by comparing the energy spectrum of all sets of sequences for five arbitrarily selected proteins from protein data bank (PDB) as evaluated from three different inter-residue pair-wise potential energy matrices. Further, continuous optimization technique has been applied on the designed sequences to obtain the lower energy bound in sequence space. The presented method enables the computation of a lower bound as well as a tight upper bound for the energy of a given conformation in sequence space. Additionally, the similarity between naturally occurring sequences and the designed sequences were examined by a pairwise sequence alignment (using BLAST program) with the non-redundant protein sequence database. The results obtained from this analysis show a good signature of similarity between existing protein and design sequences. To summarize, it is inferred that the proteins show a trend towards minimization of energy in the sequence space but do not seem to adopt the global energy-minimizing sequence. The likely reason for this could be that the protein sequence-structures are co-optimized with its surrounding environment during evolution. Also, it is likely that the existing energy matrices need to be further refined to represent the inter-residue interactions in the context of the protein environment (Jha *et al.*, 2009).

The development of pair-wise interaction energy matrices by considering the environment of amino acid in the protein is the second part of the thesis and is described in chapters 6 and 7. Here, we describe the importance of protein internal environment and its effect on the amino acid interaction preferences. In chapter 6, we considered a non-redundant dataset of globular proteins and defined two

different types of environment for amino acids in proteins. The first one is the contact based environment; in which the number of non-covalent contacts made by amino acid with their neighbors is used for classification of environments; and the second type in which the secondary structure of the amino acids is defined as an environmental parameter. Amino acids have been divided into different environments and the knowledge-based scoring matrices of various sizes based on the number of defined environments are generated. These scoring matrices are validated on two sets of decoy structures. This analysis reveal the fact that the optimal information is approximately encoded in a 60 x 60 matrix describing the 20 types of amino acids in three distinct secondary structures (helix, beta strand, and loop) (Jha *et al.*, 2010).

Chapter 7 has a brief discussion on the membrane proteins which are in a completely different environment than the soluble proteins. Here, we selected the non-redundant dataset of trans-membrane alpha-helical proteins to study the effect of internal environment with the aim of computation of knowledge-based potential energy matrix exclusively for membrane proteins. In this regard, we explored three different types of environments for the membrane spanning region of the proteins dependent on the location and the orientation of trans-membrane alpha-helices along the lipid bilayer. The scoring matrices of different sizes have been developed for the respective environments. The resultant scoring matrices are tested on a set of random sequences for five different membrane proteins that are not in the dataset. The z-score for the selected proteins is always better for scoring matrices obtained from membrane protein dataset compared to the matrices computed from globular protein dataset. In spite of few structurally solved membrane proteins, we are able to capture the differences between globular and membrane proteins. The poor z-score for membrane spanning region

of proteins from potential matrices of globular proteins proves the need of the separate scoring matrix for membrane proteins and we are able to provide one such scoring function, which performs reasonably well on membrane proteins. We also computed the contact-based environment dependent hyrdophobicity scale for amino acids in globular and membrane proteins. The results from these analyses can be used in modeling, structure prediction and docking studies.

A detailed study of helix-helix packing in helical membrane proteins is discussed in Chapter 8. Significant inter-helical interactions, between the side-chain atoms of amino acids, are considered as contacts by weighing the number of atom-atom contacts. Backbone geometry of helical residues is defined in terms of local coordinate axes, by using the C_α coordinates of amino acids. The dataset of helices has been classified into set of parallel, anti-parallel, and perpendicular pairs. The combination of contact criteria, local backbone geometry and classification of pair of helices helped in the systematic and quantitative analysis of inter-helical interactions. A single parameter (defined as alpha), which is derived from the parameters representing the mutual orientation of local coordinate axes, is able to accurately capture the details of helix-helix packing. For example, a specific range of alpha values is preferred for interactions among the anti-parallel helices. In summary, the packing in alpha-helical membrane proteins, which is systematically and rigorously investigated in this chapter, may prove to be useful in the modeling of helical membrane proteins (Jha and Vishveshwara, 2009).

Chapter 9 concludes the thesis with potential applications of the work described in it. In this thesis, we presented a computationally efficient method to design the energetically minimized sequences which can fold into a functionally

better protein than the existing one. In future, topology-based weighted graph may be applied for structure-prediction. The scheme presented here may also be included in the existing structure-prediction methods. The ranking scheme can tested against biological observations such as correlated mutations and changes in structure/function in specific proteins. The scoring matrices presented here may be applied for structure-prediction and modeling, particularly the matrix for membrane proteins may find application in that area, since this is the only matrix reported so far. Further, the environment-dependent hydrophobicity values presented here may also have application in screening ligands for protein docking.

References

1) Chothia, C., 1992. Proteins. One thousand families for the molecular biologist. Nature 357, 543-4.
2) Li, X., and Liang, J., Knowledge-Based Energy Functions for Computational Studies of Proteins, in: Xu, Y., *et al.*, Eds.), Computational Methods for Protein Structure Prediction and Modeling, Springer New York 2007, pp. 71-123.
3) Jha, A.N., and Vishveshwara, S., 2009. Inter-helical interactions in membrane proteins: analysis based on the local backbone geometry and the side chain interactions. J Biomol Struct Dyn 26, 719-29.
4) Jha, A.N., Ananthasuresh, G.K., and Vishveshwara, S., 2007. Protein sequence design based on the topology of the native state structure. J Theor Biol 248, 81-90.
5) Jha, A.N., Ananthasuresh, G.K., and Vishveshwara, S., 2009. A search for energy minimized sequences of proteins. PLoS One 4, e6684.
6) Jha, A.N., Vishveshwara, S., and Banavar, J.R., 2010. Amino acid interaction preferences in proteins. Protein Sci 19, 603-16.
7) Luthra, A., Jha, A.N., Ananthasuresh, G.K., and Vishveswara, S., 2007. A method for computing the inter-residue interaction potentials for reduced amino acid alphabet. J Biosci 32, 883-9.